Predicting Poverty levels in South Africa

Jessica van der Berg[a]

[a]*Stellenbosch University, South Africa*

Abstract

Abstract to be written here. The abstract should not be too long and should provide the reader with a good understanding what you are writing about. Academic papers are not like novels where you keep the reader in suspense. To be effective in getting others to read your paper, be as open and concise about your findings here as possible. Ideally, upon reading your abstract, the reader should feel he / she must read your paper in entirety.

Keywords: Multivariate GARCH, Kalman Filter, Copula

JEL classification L250, L100

1. Introduction

Since apartheid ended in 1994, the South African Government has committed a significant number of resources to distribute grants effectively and efficiently to the poor and the vulnerable. However, South Africa has remained a country with extremely high levels of inequality and poverty for an upper middle-income country Stats (2011). Previous literature has shown that poverty levels are higher in rural areas than in urban areas. This is largely due to rural household not having access to the same employment opportunities has urban households. South Africa is a country with high levels of unemployment and extremely low wages. This implies that individuals that are economically active can still fall below the upper or lower bound poverty line due to the low wages that they receive Leibbrandt, Woolard, Finn & Argent (2010).

Poverty remains unacceptably high for a country of South Africa's economic status and remains closely associated with race. Thus, poverty reduction remains one of the key economic goals. Poverty in money terms has declined markedly since apartheid ended in 1994. This was made possible by the expansion of the social grant system. However, accurately targeting social welfare programs can be challenging given that income data is often incorrect Yu & Van der Berg (2017). To overcome this problem, households are subject to a proxy means test (PMT) to identify whether a household qualifies for social assistant. This essay will try to identify new methods to identify households that

Email address: 20190565@sun.ac.za (Jessica van der Berg)

are below the food poverty line, lower-bound poverty line and upper bound poverty line using machine learning techniques.

Structure ... ‘

2. Literature Review

Write a short literature review here on the background of SOuth Africa - max one page

3. Data

The data used for predicting poverty level was exacted from the General Household Survey (GHS) 2018, which is a survey completed annually by Statistics SA to measure the living circumstances of households in South Africa. They survey includes household and individual characteristics. After taking out all of the NA values and taking out households who did not know or answer the relevant questions, the dataset consists of 14 546 entries.

Since the GHS consist of household survey, total income per household is reported. To calculate the average monthly income per individual within each household, I first need to calculate the total number of adults within each household. This is done by taking the difference between household size and the number of children under the age of 17. I then take the total monthly income and divide it by the total number of adults to get the average monthly income per individual in each household. This income information is then used to divided the data into four groups.

The first group is individuals whose income fall below the food poverty line. In 2018, the food poverty line was R547 per month. The food poverty line is also referred to as the extreme poverty line as it refers to the absolute minimum amount an individual will need to be able to afford the minimum energy intake for survival. The second group consist of individuals whose monthly income falls between the food poverty line and the lower-bound poverty line (R785). The lower bound poverty line is the sum of the food poverty line and and minimum amount for non-food items. The third group consist of individuals between the lower-bound poverty line and the upper-bound poverty line (R1183). The final group consist of individuals whose monthly income is above the upper-bound poverty line, which I refer to as non-vulnerable individuals. Figure 1 graphically displays the process described above in the format of a decision tree, where 1 represents the food poverty line, 2 the lower-bound poverty line, 3 the upper-bound poverty line and 4 the non-vulnerable.

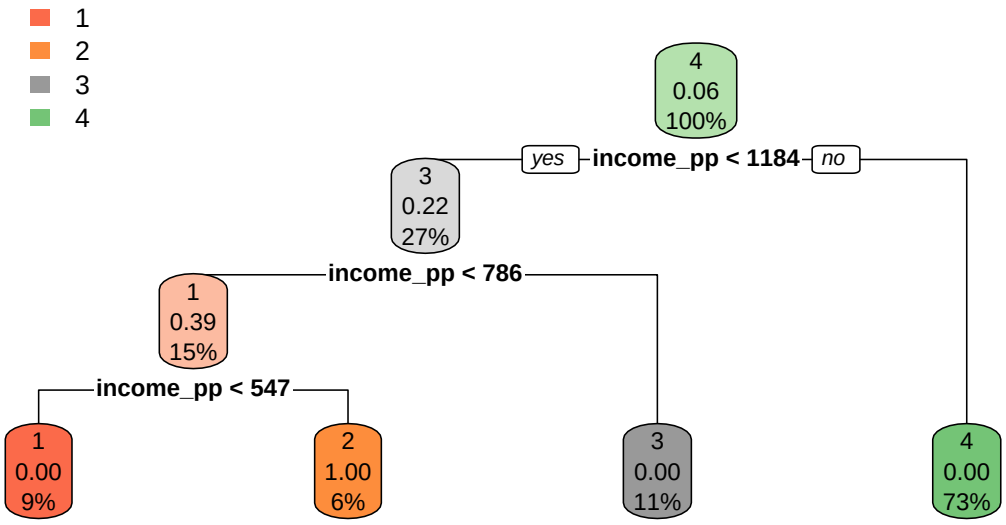

Figure 3.1: Decision tree for poverty levels

After all data cleaning was done, 73 percent of households were non-vulnerable, 11 percent were between the upper and lower-bound poverty line, 6 percent between the lower and food poverty line, and 9 percent fell below the food poverty line. To evaluate the performance for the machine learning techniques that I will implement, I randomly split the data into two subsets using a 70:30 ratio. The training dataset will consist of 70 percent of the original dataset, while the test dataset will consist of the remaining 30 percent. Figure 2 below shows the number of household per poverty level for the training dataset.

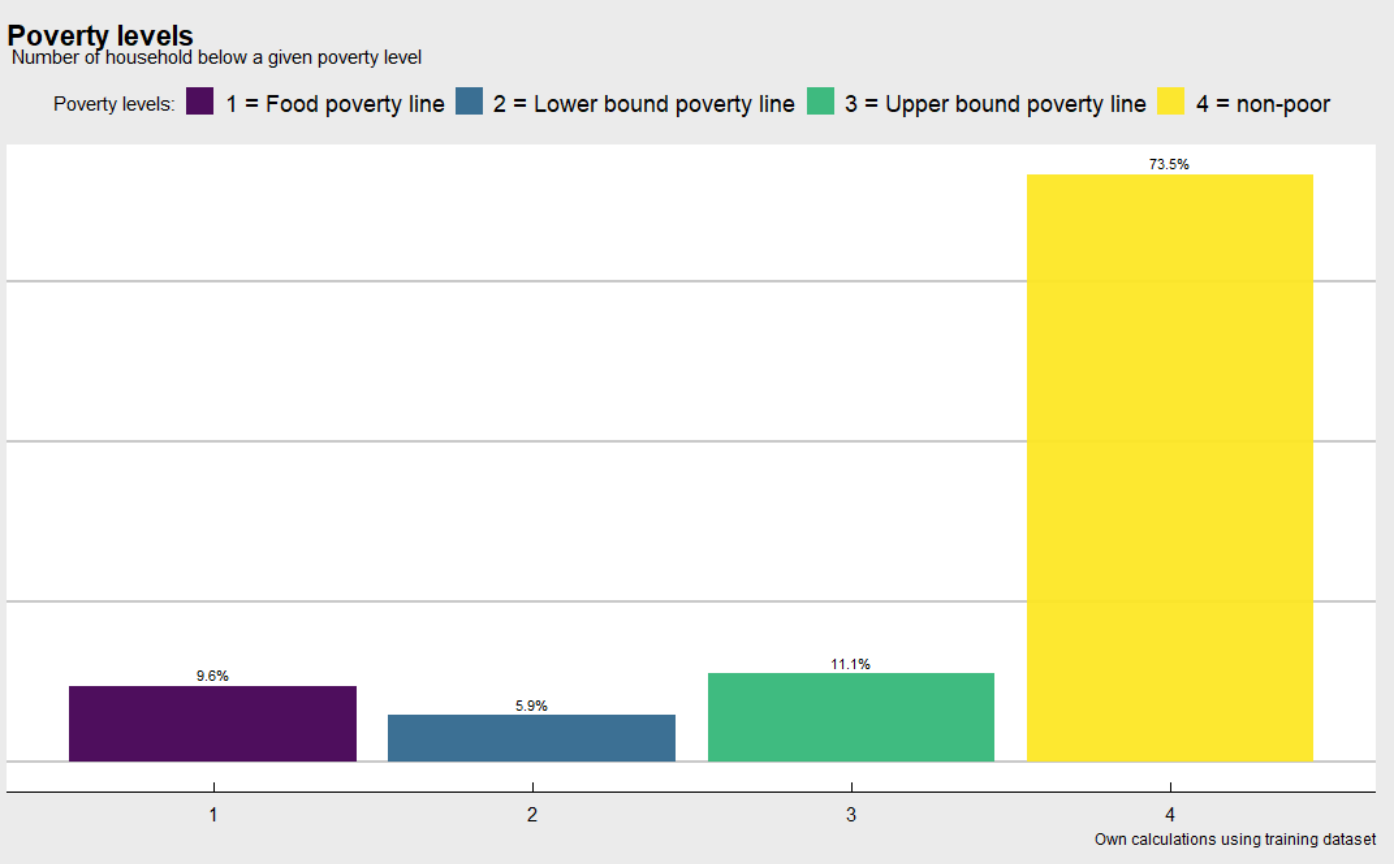

Figure 3.2: Number of household per poverty level

Figure 2 shows that in our training dataset, 73.5 percent of households are non-vulnerable and 9.6 percent of households monthly income falls below the food poverty line. This implies that I have an extremely imbalanced dataset. This is important as it will have an effect on the machine learning techniques that I implement later on.

For some descriptive statistics, I analyze some of the variable in more details. From the density plots in figure 3.3, I can see that households whose average income per person falls below the upper-bound poverty line (green line) tend to have slightly older head of household (head_age) then those household belonging to other poverty levels.

A household size (hholdsz) of 1-2 individuals tend have a larger probability to be non-vulnerable (yellow line), where larger household size of 3-5 individuals then have a larger probability of falling below the food poverty line.

Total expenditure (Q814Exp) is higher for the non-vulnerable households, which makes sense as they receive a higher income.

The total number of rooms (Q55TotRm) seem to be distributed similarly, with non-vulnerable households having a slightly smaller distribution. This comes as no suprise as these non-vulnerable households tend to have a smaller household size, implying that they need less bedrooms.

Non-vulnerable household also have much larger property valuation (Q58Val), which could imply that they have a higher standard of living.

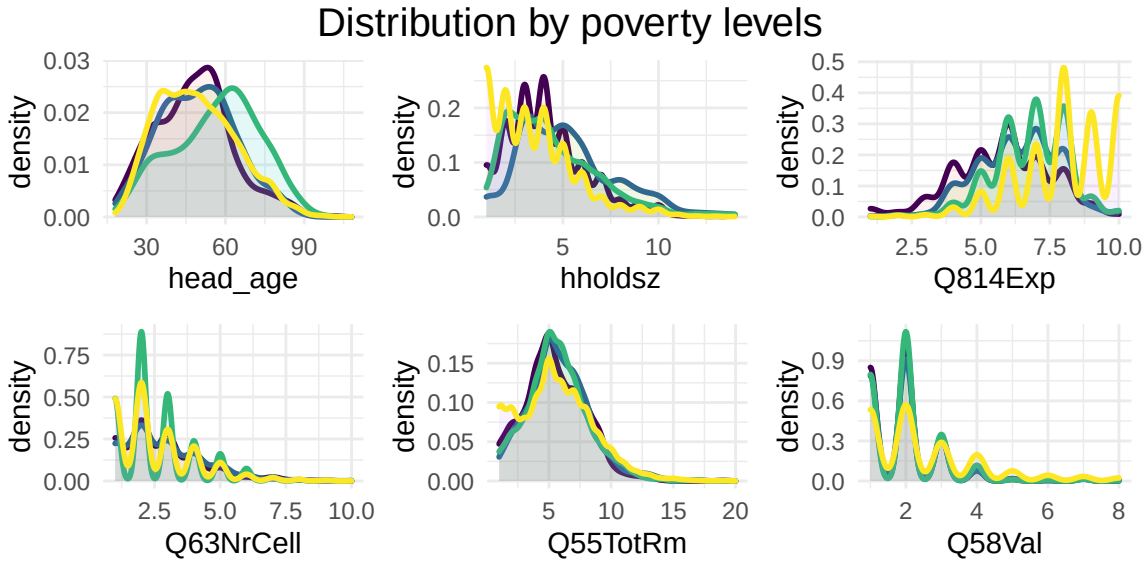

Figure 3.3: Distribution of certain variables by poverty level

Before we start building models we can also analysis which variables are correlated with our dependent variable (which is is poverty level). Figure 3.4 displays the top 10 variables that are correlated with the dependent variable. Our top four that are positively correlated are *econact hh*, which is a binary variable indicating whether or not the household is economically active, *Q814Exp*, which is a numeric variable display total household expenditure, *totmhinc*, which is a numeric variable displaying total household income, and *Q89aGrant*, which is a binary variable indicating whether the household receives a grant or not.

variables that are negatively correlated with the dependent variable is the sex of the head of the household (*head sex*), whether or not the household owns a washing machine (*Q821WashM*), the household size (*hholdsz*), the amount spent on rent or mortgage of property per month (*Q57Rent*) and the number of adults in the household (*adults*).

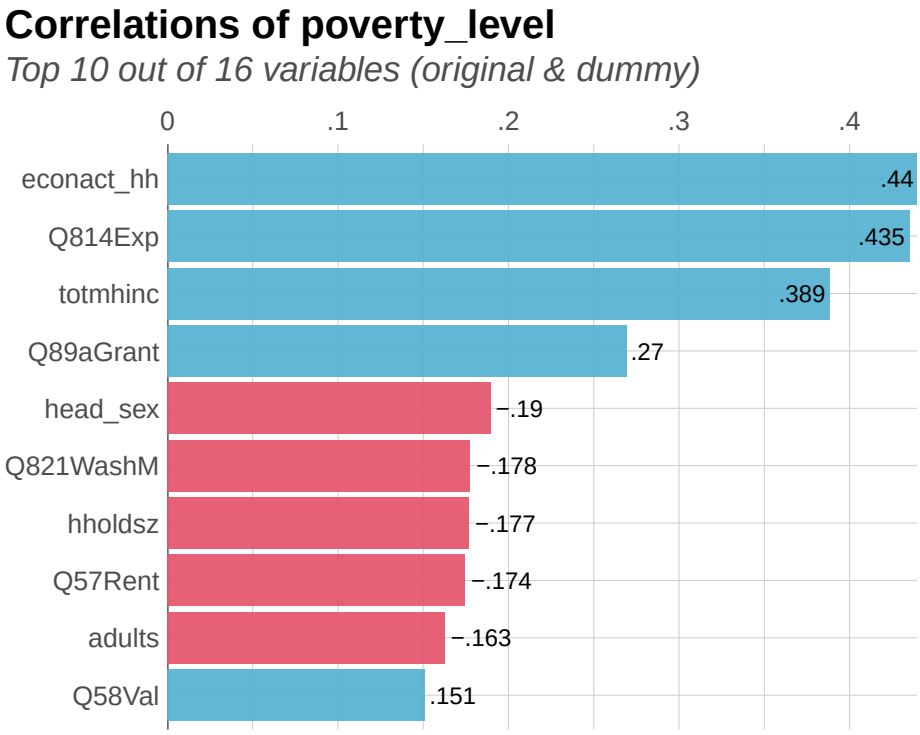

Figure 3.4: Correlation with dependent variable

4. Methodology

4.1. Analysing Classification Models

To see which model predicts poverty most accurately and efficiently, I assess the performance of x supervised classification models using the caret package. The dataset is extremely unbalanced; therefore, the defaults parameter will be Kappa to order to improve the performance of the models. The Kappa metric compares the observed accuracy with the expected accuracy. It also accounts for random chance which implies that it makes the model more accurate that simply using an Accuracy as a metric. The Kappa matric is calculated using the formula;

$$k = \frac{p_0 - p_e}{1 - p_e}$$

Where p_0 represents the overall accuracy of the model and p_e represents that measure of the agreement between the predictions and actual class value of the model. Therefore, Kappa attempts to account for evaluation bias by considering the correct classification by a random guess.

The table below briefly discusses the seven different classification models that I compared to determine which model is the most accurate to predict poverty Boehmke & Greenwell (2019).

Model	Reference Name	Description
Multinomial Logistic Regression	multinom	Makes use of the maximum likelihood estimation to evaluate the probability of a categorical relationship
Linear Discriminant Analysis	lda	Used to find linear combinations of separates multiple classes of features, representing the dependent variable as a linear combination of other features
Naive Bayes	naive bayes	Using Bayes Theorem, this model applies posterior probability to the categorization, making the uneducated assumption that the predictors are independent
Linear Support Vector Machine	svmLinear	The model creates a line that separates data into classes
K-Nearest Neighbor	knn	Each observation in the dataset is predicted based on its similarity to other observations
Recursive Partitioning	rpart	Builds models using a general structure which consist of a two-stage procedure and then final presenting the model as a binary tree
Ranger	ranger	An updated and fast implementation of random forest for big data

4.1.1. Evaluation

To ensure that the correct model is chosen, the accuracy of each model is evaluated and compared. However, since the dataset is unbalance, I also analyze the F measure, also known as the F_1 score, of each model. The F_1 score communicates the average between the precision and the recall of each model. A perfect model has an F_1 score equal to 1, therefore models with a higher F_1 score is preferred over models with a lower F_1 score. The formula for the F_1 score is given below;

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

The extremely unbalanced data will affect the results of the F_1 score. Therefore, it is also informative to evaluate the macro F_1 score and the weighted F_1 score. The macro F_1 is not affected by unbalanced data and is equal to the average of the F_1 score and is commonly used when there are multiple levels or classes. It gives the same importance to each poverty level. A higher macro F_1 score is preferred to a lower one. The formula for the macro F_1 score if given below;

$$Macro\, F_1\, score = \frac{1}{N}\sum\nolimits_{i=0}^{N} F_1$$

Where N is the number of different poverty levels and i is the levels index. The drawback of the macro F_1 measure is that is gives equal weight to all poverty levels, which implies that it over emphasis the under-represented poverty levels. The weighted average F_1 score is similar to the macro F_1 score, but here the F_1 score is weighed according of the number of households from the specific poverty level, which emphasis poverty levels according to size of each poverty level. The formula for the weighted F_1 score is given below;

$$Weighted\, F_1\, score = \frac{n_i \sum_{i=1}^{k} F_1}{\sum_{i=n}^{k} n_i}$$

5. Results and Discussion

5.1. Classification Models

The table below shows the different metrics to make comparing different models easier. The first feature that is observed is that the models vary drastically in time. The *lda*, *rpart* and *naive bayes* models are extremely fast whereas *multinom* and *ranger* take relatively long to run. Furthermore, *ranger* scores the best for accuracy and F_1 measures.

Models	**Time**	**Accuracy**	**Macro F_1**	**weighted F_1**
multinom	37.17	0.9935839	0.9820473	0.9935798
lda	0.81	0.7774977	0.4561088	0.7409658
naive bayes	1.67	0.8819890	0.8138291	0.8862441
svmLinear	9.91	0.9660862	0.9134729	0.9661192
knn	4.81	0.9869386	0.9714937	0.9876236
rpart	1.08	0.9372136	0.9155884	0.9137702
ranger	41.92	1.00000000	1.00000000	1.00000000

Analyzing the accuracy score, Figure 5.1 shows that *ranger* has the highest accuracy, followed closely by *multinom*. However, these two models also take the longest to run. *knn* and \emph{rpart) are relatively fast and thet have a high level of accuracy. For the fastest running models, decision trees (*rpart*) are the most accurate. Figure 5.1 also implies that there is some trade-off between accuracy and computational time.

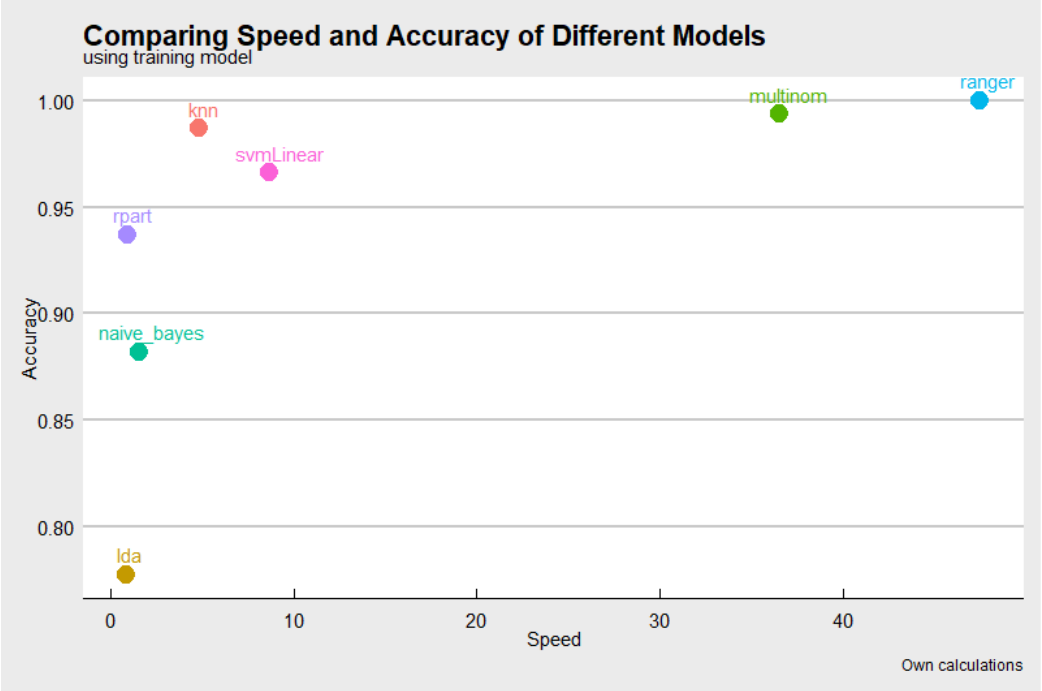

Figure 5.1: Speed versus Accuracy

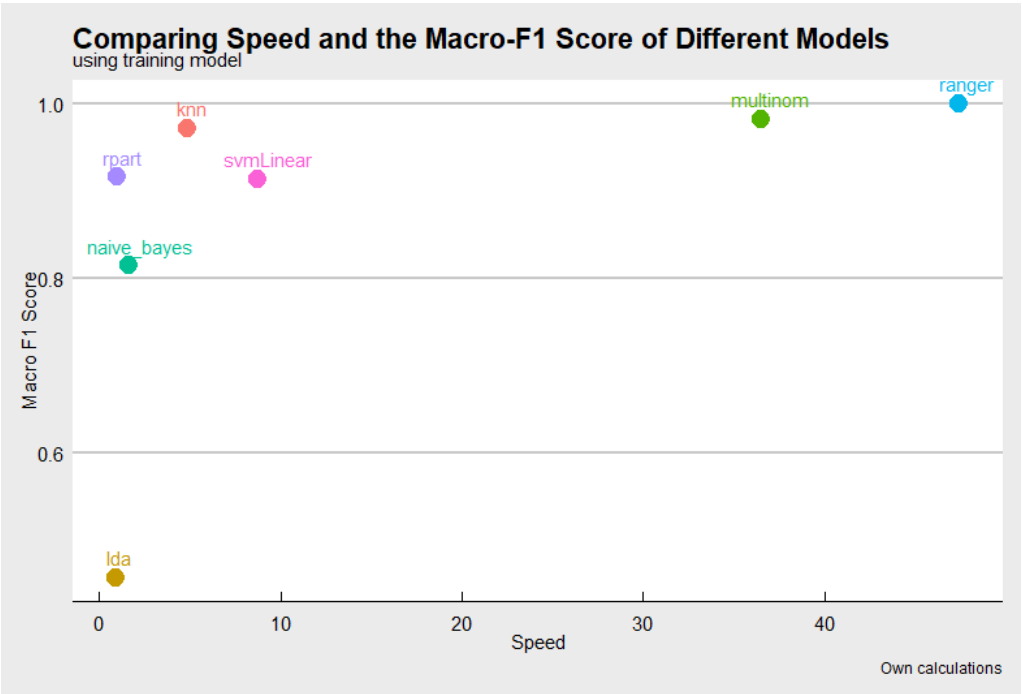

Figure 5.2: Speed versus Macro F1

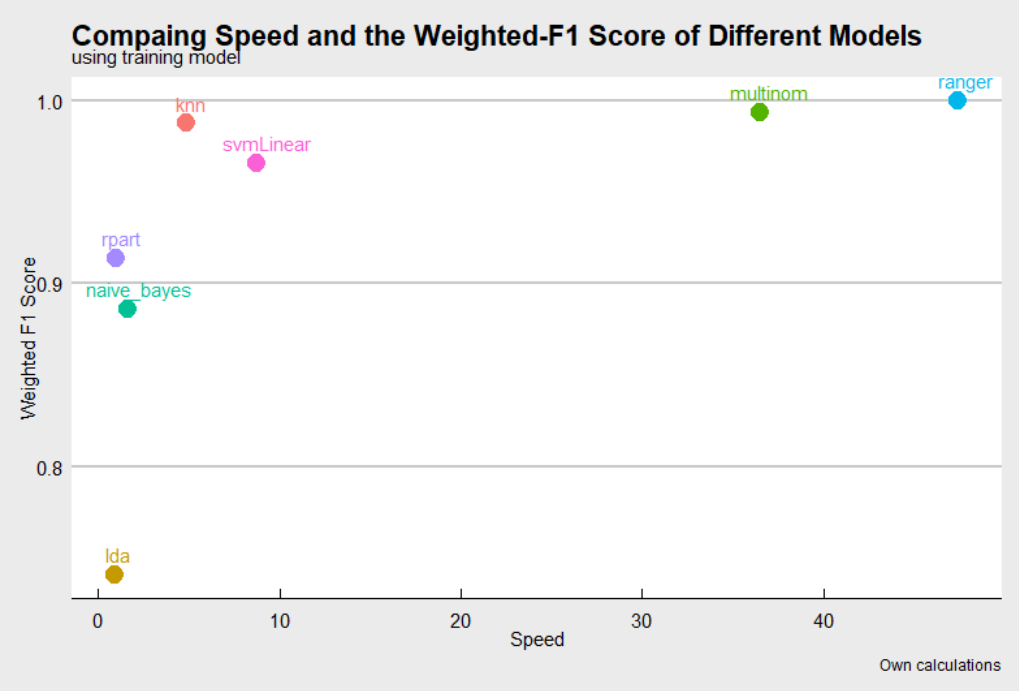

Figure 5.3: Speed versus Weighted F1

References

10 Boehmke, B. & Greenwell, B.M. 2019. *Hands-on machine learning with r*. CRC Press.

Leibbrandt, M., Woolard, I., Finn, A. & Argent, J. 2010. Trends in south african income distribution and poverty since the fall of apartheid.

Stats, S. 2011. Social profile of vulnerable groups in south africa 2002-2010. *Pretoria: Government Printer*.

Yu, D. & Van der Berg, S. 2017. South african poverty: The current situation and trends since the transition to democracy.

Appendix

Appendix A

Some appendix information here

Appendix B